Adaptive Wavelet Eye-Gaze Based Video Compression

Mohsen Farid[1], Fatih Kurugollu, Fionn Murtagh
School of Computer Science, Queen's University, Belfast, UK

ABSTRACT

We present a novel approach to compression of video frames based on the foveation behavior of the human visual system (HSV). Eye fixations on a video frame, as depicted by eye-gaze trace data, define an imaginary region of interest. The perceived resolution of the frame by the human eye depends totally on this eye-gaze (fixation) point. The resolution, then, decreases dramatically with the distance from the fovea. This behavior of the HSV has gained interest in the image and video processing area recently especially in compression of images or video frames. We present an approach where eye-gaze trace data are integral to the compression process which has demonstrated its usefulness in yielding high compression performance. We partition a video frame into three regions: the inner-most includes a point of eye-gaze for which we apply lossless compression; an outer region which encompasses the first and for which we apply visually lossless (near-lossless) compression, and finally an outmost region where lossy compression is applied. Because of its low computational complexity, we use the Haar wavelet transform. Preliminary results are promising and show improvement over other methods which are mainly full frame based.

Keywords: Video compression, wavelet transform, foveation, eye-tracking, ASL-504.

1. INTRODUCTION

Video compression is one of the key technologies for digital media. It is used to store and transmit the video frames efficiently. The main concern in video coding is to compress the video frames as much as possible without significant degradation of the visual quality. This objective can be carried out by using four kinds of redundancies in the video frames: spatial, temporal, spectral and psychovisual[1]. While the first three redundancies inherently rest in the data in the video frame, the last one stems from the characteristics of the human visual system (HSV). These characteristics are primarily the luminance, frequency and texture masking. These have already been investigated in compression systems and have yielded not only better compression ratios but also better visual quality.[2]

Another interesting property of the HSV, which has gained more attention in the research community lately, is the space-variance nature of the human eye. The retinal tissue consists primarily of two classes of photoreceptors: the cones and the rods. Cone photoreceptors respond to daylight stimuli whereas rods respond to low light intensities. Visual acuity increases as the visual stimuli come closer to the fovea, as indicated by the increase of cone cells density. One may view this from an engineering perspective whereas the spatial resolution is proportional to the cone cell density which also decreases as the distance of the stimulus from the fovea increases. Resolution reaches its peak value at high luminance values at a zero value of visual angle from the fovea. At very low luminance values stimuli cannot not be in the fovea. They could be seen further out from the fovea. Furthermore, as they move closer to the fovea no noticeable improvement in the visual acuity is observed. One may conclude that high frequencies stimulating areas farther away from the fovea do not contribute highly to the overall image perceived by the eye. Furthermore, one may, then, devise a compression mechanism that lessens the effect of high frequencies away from the fovea thereby achieving higher compression performance.[3, 14, 15]

Conventional video compression techniques such as MPEG-I/II use digital representation of the original video data in which a uniform sampling in space (and time) is used for the conversion and hence they use space-invariant sampling. One may additionally assert that the space-variant sampling property of the HSV may be used to improve the performance of current coding systems by eliminating the redundancy inherent in pixels away from point of gaze as given by eye trace data.[3]

[1] Corresponding author: m.farid@qub.ac.uk, +44 28 90 26 46 21, School of Computer Science, Queen's University Belfast, Belfast, BT7 1NN, United Kingdom.

There has been growing interest in foveation based image and video compression. We envisage a system of news cast or teleconferencing where normally on person is the focus of attention during a video clip as a possible application of foveated compression scheme. The early attempts using foveation property of the HSV were based on grouping local pixels and averaging and mapping them into superpixels. The sizes of the superpixels are determined by the spatial density of the photoreceptor cells[4]. The foveated frames can be obtained by a filter process. A foveated filter bank which contains low-pass filters having variable cutoff frequencies is applied to the frames. This technique has been used in MPEG-II/H.263 video coding applications[5]. Chang and Yap have introduced a wavelet based foveation method. They used the wavelet transform to obtain foveated images by means of a non-uniform weighting model for the wavelet transform[6].

Wang and Bovik have recently carried out wavelet based foveation work in which they did not use wavelet filters to obtain foveated images. Instead, they used a foveation-based error sensitivity model in the wavelet transform domain. This model has been exploited to weight wavelet coefficients according to eye gaze point. The weighted coefficients were coded using a modified SPIHT algorithm which benefits from the foveation-based error sensitivity model. The model depends on the cuttoff frequency in the sense that any higher frequency component beyond it is invisible (The cutoff frequency depends on the distance from the fixation point and the distance from the screen), the display resolution, and, finally, the visual importance of the wavelet coefficients at different subbands[3]. Although this study provides a sound mathematical framework for foveation based wavelet image coding, the algorithm results in high computational complexity which may negatively affect the use of the algorithm in real-time applications.

In this paper, we propose a new adaptive wavelet technique to overcome this problem. To implement the wavelet transform in real time, a video frame is divided into non-overlapping subblocks. To achieve lossless or near-lossless coding, a reversible integer wavelet transform is used. For each subblock, the wavelet coefficients are coded adaptively according to the location of the current eye-gaze point. Eye-gaze trace data are collected using the ASL 504 Pan/Tilt Eye-Tracker. The organization of the paper is as follows: in Section 2, we describe the coding approach that we use. Section 3 discusses the reversible integer wavelet transform that is used in the system. The results are presented and discussed in Section 4, and the concluding remarks are in Section 5.

2. REAL-TIME VIDEO CODING SYSTEM with EYE TRACKER

The block diagram and an image of the system are depicted in Figure 1 and 2, respectively. While the observer watches the video clip, the video frames are captured and added to the eye gaze coordinates stream as in Figure 4. Eye gaze points are represented in the figure by the intersection point of the vertical and horizontal lines of the black cross-hair.[16,17] The video clip is encoded in a frame based fashion which means that each frame is processed independently, similar to M-JPEG.

To fulfill real time constraints in a foveation based coding, we adopt a sub-block coding strategy. This means that the wavelet transform is implemented on sub-blocks rather than the entire frame. The sub-block coding concept has been widely used in image/video coding systems for different purposes. In the JPEG standard, for instance, 8x8 sub-blocks are used to calculate and code DCT coefficients. In MPEG-I/II video coding standards, the same block size is also used for the motion compensation step[1]. Recently, the JPEG2000 image coding standard, which is based on wavelet transform, uses a tiling scheme, which refers to partitioning of the image into rectangular non-overlapping blocks.[7] The blocking scheme does not only improve the real time performance but also reduces the memory requirements for hardware solutions.

In the proposed system, the sub-blocks are subjected to a reversible integer wavelet transform. For color frames, each color band is coded as a separate sub-block successively. A Haar wavelet transform is used for its low computational complexity. We also use a different wavelet coefficient coding scheme for each sub-block depending on the current eye-gaze point. We defined three types of sub-blocks: 1) lossless sub-blocks. 2) near-lossless sub-blocks (visually lossless) and 3) lossy sub-blocks. The lossless sub-blocks (marked as ‘A’ in Figure 1) reside around the current eye-gaze point, which is given by the eye tracker. Sub-blocks (A) are coded using wavelet coefficients in a lossless manner since they are centered in the foveation area. Near-lossless blocks (marked as ‘B’ in Figure 1) are layered around the lossless sub-blocks. While there is a little loss of image quality, mathematically speaking, visual frame image quality is maintained.

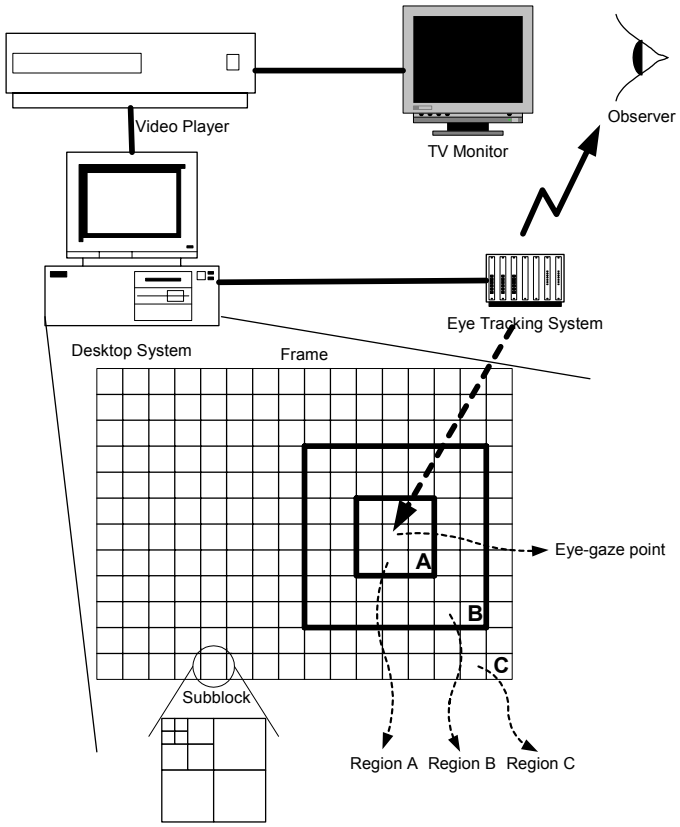


Figure 1: Block diagram of the system.

Figure 2. Eye is tracked using an ASL 504 Tilt/Pan infra red camera which is connected to a host computer through a control box. Video signal and eye trace data are combined into another workstation (partially obscured by the computer monitor) for processing.

Frame image sub-blocks outside of region B, region C in the Figure, are coded in a lossy manner since they are considerably further from the point of eye-gaze.

The coding of wavelet coefficients are carried out by a modified EZW algorithm described by Welstead[8]. The algorithm is derived from the well-known embedded zerotree wavelet (EZW) coder algorithm[9]. This algorithm

exploits the self similarity of the wavelet transform across different scales. It sorts the coefficients, including the LL subband, by magnitude and transmits the bit planes according to their order. Welstead used the same algorithm by excluding LL subband from the process. LL subband are transmitted to the receiver without compression. Since LL subband is a close approximation of the image (sub-blocks in our case), retaining this subband gives us better reconstruction of the image without sacrificing the low bit rate. The other subbands are coded in the same fashion in EZW. If we proceed as deeply as the maximum number of bit planes, we can obtain totally lossless reconstruction in the receiver since the used wavelet transform is reversible. This scheme is used in lossless subblocks.

For near-lossless sub-blocks, the EZW algorithm progresses until it reaches half size of number of bit planes. This results in some loss in detail of the sub-block, but using the original LL subband in the reconstruction makes this loss tolerable. Since the lossy sub-blocks are far from the foveation area, the details in these sub-blocks cannot be perceived by the subject. Therefore, only LL subband coefficients are transmitted to the receiver. This scheme also speeds up the process since it does not waste time to decode the coefficients of the other subbands.

An example of the coding result can be seen in Figure 3. This figure shows a frame from the 'Akiyo' video clip. Figure 3.b represents the coding scheme by partitioning the frame into subblocks. The eye-gaze point is fixed on the face of the speaker. Therefore, the lossless blocks are centered at this point, which are shown as a shaded area. The near lossless sub-blocks surround the area occupied by lossless sub-blocks. The area outside of this near lossless region is designated as the lossy sub-block region. The decoded frame is shown in Figure 3.c. Frame image loss can be seen in the lossy sub-blocks region.

3. REVERSIBLE WAVELET TRANSFORM

The wavelet transform has been widely used in image/video compression with great success. Furthermore, tt has been accepted as core processing tool in the JPEG2000 image compression standard[7]. The benefit of the wavelet transform lies behind its good energy compaction property as well as its progressive nature of producing embedded bit-stream. However, wavelet transforms have their limitations. The filter coefficients are normally real numbers. While image pixels may be represented by integer values, transform coefficients would be real. This therefore could lead to lossy coding of the image.

A solution to this problem is integer wavelet transform,[10,11] which have successfully been used in image compression.[12,13] Using the integer transform, the wavelet coefficients can exactly be reversed. In such cases it is termed as "reversible" wavelet transform.[12] Basically, reversible wavelet transforms are derived from their linear wavelet transform counterparts, rounding filter outputs to the nearest integer. It exploits the redundancy in the summation and the difference of two integers as depicted especially in the Haar wavelet transform. The sum and the difference, in the Haar wavelet transform, have same least significant bit. This redundancy can be eliminated by dividing the summation by 2 (shift right by 1)[12].

Consider the Haar wavelet transform as follows:

$$\begin{aligned} r(i) &= \frac{x(i)+x(i+1)}{2} \\ d(i) &= \frac{x(i)-x(i+1)}{2} \end{aligned} \tag{1}$$

where x(i), r(i) and d(i) denote input signal, low pass filter output and high pass filter output, respectively. It is evident that the output would be half integer when either of x(i) or x(i+1) is odd.

Zandi et al has shown that an efficient reversible (integer) version of the Haar transform can be defined as follows[12]:

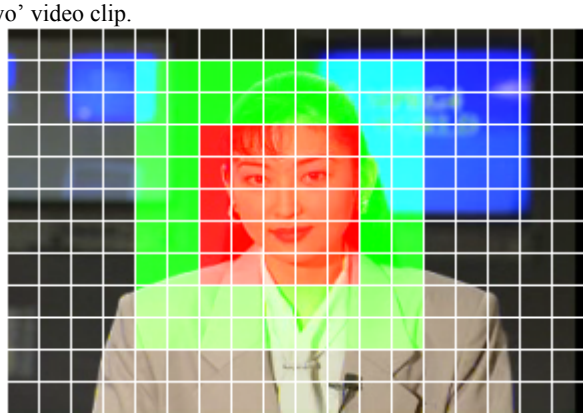
$$\begin{aligned} r(i) &= \left\lfloor \frac{x(i)+x(i+1)}{2} \right\rfloor \\ d(i) &= x(i)-x(i+1) \end{aligned} \tag{2}$$

Here the floor operation makes the transform not only integer but also reversible. Using the aforementioned redundancy, the inverse transform can be defined as follows:

$$\begin{aligned} x(i) &= r(i) + \left\lfloor \frac{d(i)+1}{2} \right\rfloor \\ x(i+1) &= r(i) - \left\lfloor \frac{d(i)}{2} \right\rfloor \end{aligned} \tag{3}$$

It can easily be shown that such a transform has a low computational complexity. The forward transform needs only two additions and one shift operation. The inverse transform, on the other hand, requires two additions, one increment and two shift operations. Moreover, these integer operations can be implemented in a parallel manner by using the MMX instruction set in the Pentium processor based PCs[11].

(a) Original frame form the 'Akiyo' video clip.



(b) Partitioning of the frame. The inner area that represents the foveation area and is coded losslessly. The ring around the inner area is coded in near-lossless manner, while the outermost sub-blocks are coded loosely.

(c) The coding result. The foveation area is exactly reconstructed while the details around this area are compromised.

Figure 3. An example of coding strategy.

4. RESULTS

We have tested our algorithm on a real video clip with eye-gaze trace data. The eye-gaze position data are added to the video frames in real time. A small set of these video frames are shown in Figure 4. The eye-gaze positions are depicted by the black cross-hair.

The reconstructed frames of 320x240 pixels are depicted in Figure 5. We have used 16x16 sub-blocks and 3 level wavelet decomposition for this adaptive foveation coding. The lossless area size consists of 5x5 sub-blocks while the near-lossless ring size is 9x9. Therefore, there exist 25 lossless subblocks, 56 near-lossless subblocks and 219 lossy subblocks. The real time performance of the algorithm is given in Table 1. In this table, the average coding times for each type of sub-blocks and the entire frame are given. The difference between the whole frame coding time and the total sub-blocks coding time is due to data management and time needed to determine the eye-gaze point. This coding times have been gauged by a 800 MHz Pentium III processor desktop PC. The C++ programming language was used to implement these algorithms. We expect considerable performance improvement when MMX technology is used and in addition when multi-thread techniques are implemented.

Subblock Type	Average coding time (ms) per frame.	Number of sub-blocks	Totals for sub-blocks (ms)
Lossless	0.84	25	21.00
Near-lossless	0.35	56	19.60
Lossy	0.24	219	52.56
Total for all sub-blocks			93.16
Entire Frame including administrative house keeping)			96.30

Table 1. Real time performance of the algorithm.

Table 1 shows the performance of the algorithm used this paper. The column “average coding time (ms)” accounts for the time in millisecond taken to compress region “A” blocks using lossless compression, region “B” blocks using near lossless scheme and the region “C” blocks using a lossy scheme. The column “Number of sub-blocks” gives the number of blocks for each of the compression scheme in each frame. The “Totals for sub-blocks” gives the total time taken for the blocks of each scheme within the one frame. The total time of 96.30 ms indicates that we can process approximately 10 video frames per seconds. On a faster 2 GHz machine we noticed almost real time performance.

The real time compression performance of video frames using the proposed sub-blocking scheme has been found to be superior to that of the entire frame scheme. As seen in Figure 6, both the real time performance and compression ratio of the proposed system is nearly four times better than the entire frame based system.

5. CONCLUSION

In this paper, we present a novel approach for foveation based video coding in real time. The system obtains the current eye-gaze point from an eye tracker and encodes it onto the frame in real-time. A sub-blocking mechanism is devised. The sub-blocks are classified and encoded as lossless, near-lossless and lossy in relation to the location of the point of gaze. To achieve a real-time solution, the Haar wavelet transform is chosen for its low computational complexity. A noticeable improvement is gained when using the sub-blocking mechanism versus a whole frame compression. The results are excellent and are expected to improve even further when better wavelet transform functions are used.

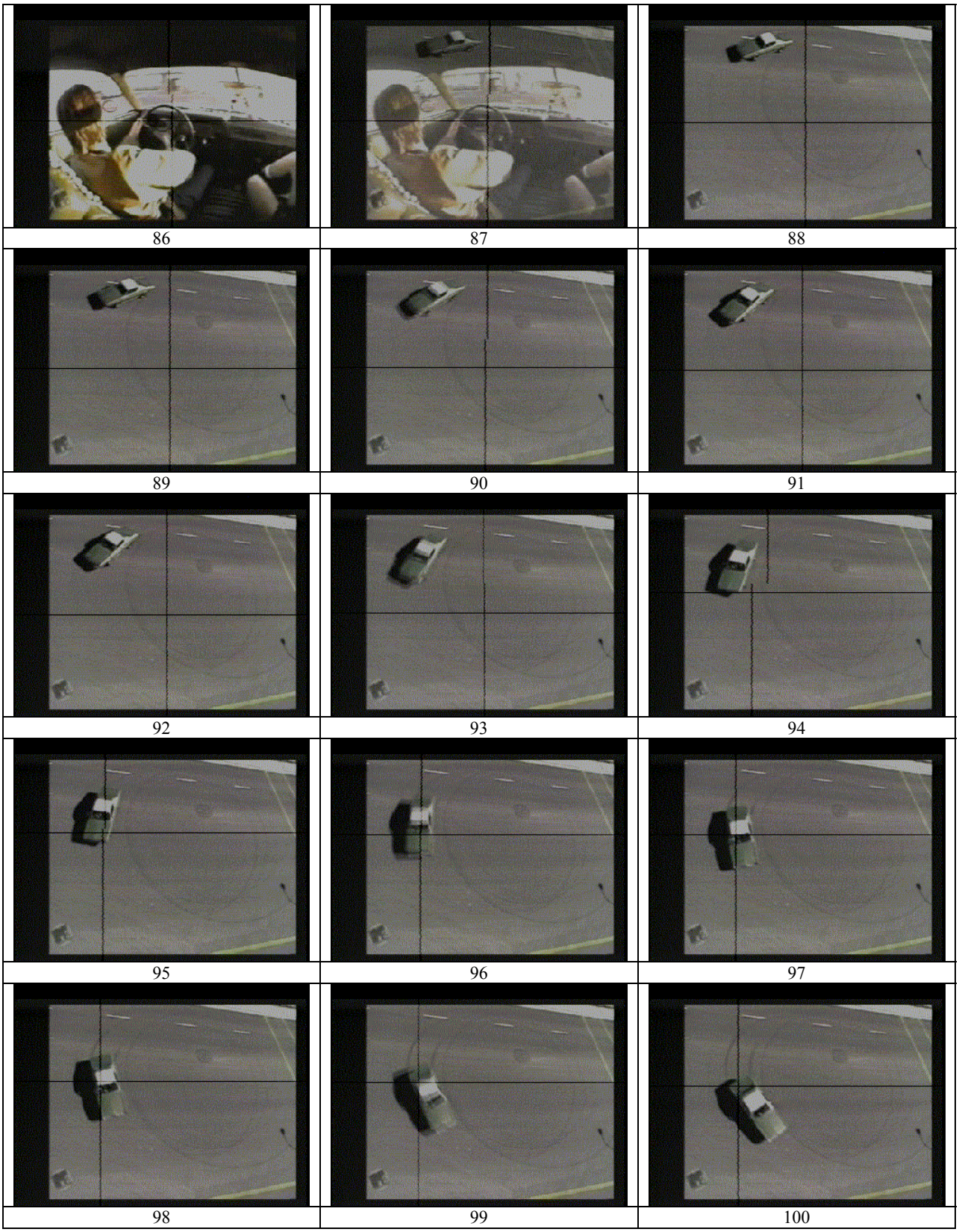


Figure 4. An example of video frames with eye gaze points.

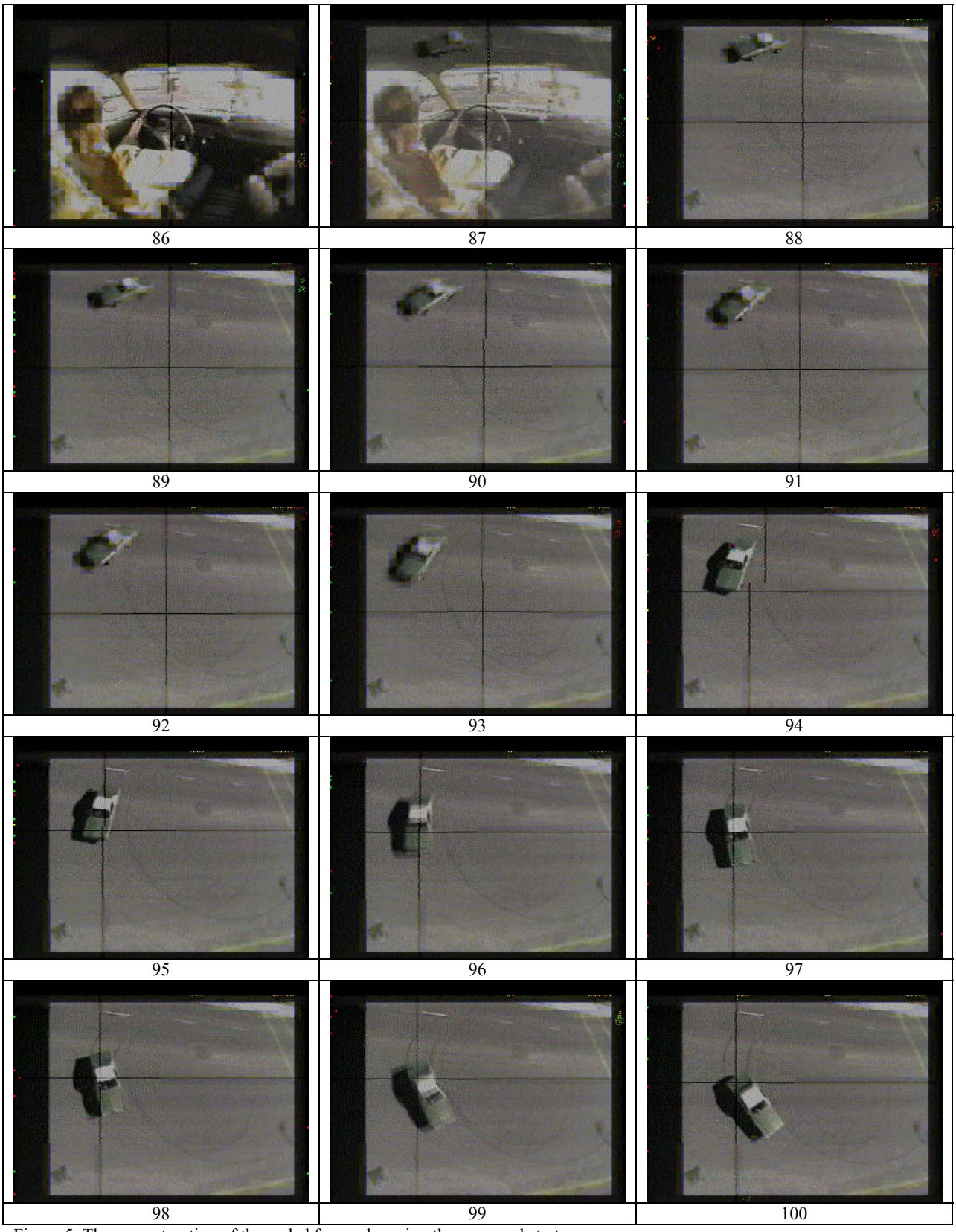
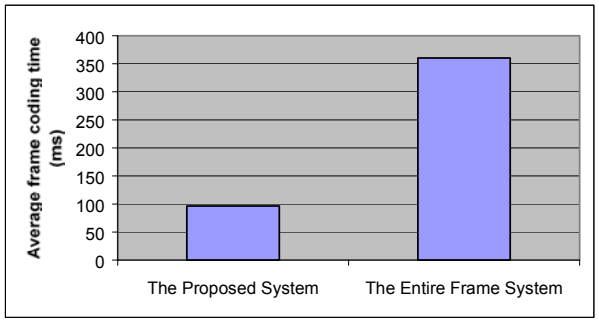
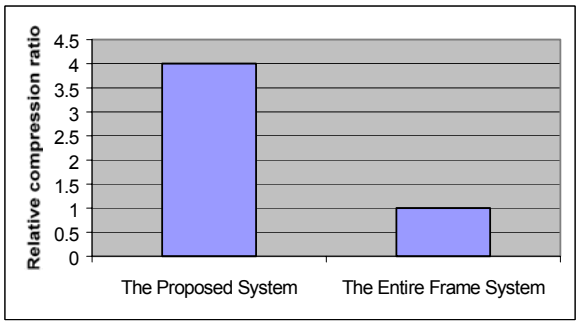


Figure 5. The reconstruction of the coded frames by using the proposed strategy.



(a) Comparison of real time performance



(b) Comparison of compression ratio

Figure 6. Comparison of the proposed algorithm with the entire frame counterpart in real time and compression ratio performances.

REFERENCES

1. A.M. Tekalp, Digital Video Processing, Prentice-Hall, New Jersey, 1995.
2. Y.Q. Shi, H. Sun, Image and Video Compression for Multimedia Engineering, CRC Press, Florida, 2000.
3. Z. Wang, A.C. Bovik, "Embedded Foveation Image Coding", IEEE Trans. on Image Processing, **10**, pp.1397-1410, 2001.
4. P. Kortum, W.S. Geisler, "Implementation of a Foveated Image Coding System for Image Bandwidth Reduction", Proc. SPIE: Human Vision and Electronic Imaging, **2657**, pp.350-360, 1996.
5. S. Lee, M.S. Pattichis, A.C. Bovik, "Rate Control for Foveated MPEG/H.263 Video", IEEE ICIP, **2**, 1998.
6. E.C. Chang, C.K. Yap, "A Wavelet Approach to Foveating Images", Proc. 13th ACM Symposium Computational Geometry, pp. 397-399, 1997.
7. A. Skodras, C. Christopoulos, T. Ebrahimi, "The JPEG 2000 Still Image Compression Standard", IEEE Signal Processing Magazine, **18**, pp. 36-58, 2001.
8. S. Welstead, Fractal and Wavelet Image Compression Techniques, SPIE Optical Engineering Press, Washington, 1999.
9. J.M. Shapiro, "Embedded Image Coding Using Zerotrees of Wavelet Coefficients", IEEE Trans. on Signal Processing, **41**, pp. 3445-3462, 1993.
10. R. Calderbank, I. Daubechies, W. Sweldens, B.L. Yeo, "Wavelet Transforms that Map Integers to Integers", Appl. Comput. Harmon. Anal., **5**, pp.332-369, 1998.
11. J. Reichel, G. Menegaz, M.J. Nadenau, M. Kunt, "Integer Wavelet Transform for Embedded Lossy to Lossless Image Compression", IEEE Trans. on Image Processing, **10**, pp. 383-392, 2001.
12. A. Zandi, J.D., Allen, E. L. Schwartz, M. Boliek, "CREW: Compression with Reversible Embedded Wavelet", Proc. Data Compression Conference, Utah, USA, pp. 212-221, 1995.
13. I. Singh, P. Agathoklis, A. Antoniou, "Lossless Compression of Color Images Using an Improved Integer-Based Nonlinear Wavelet Transform", IEEE International of Symposium on Circuits and Systems, Hong Kong, pp. 2609-2612, 1997.
14. Mountcaslte, V. B. (Ed.), Medical Physiology, 14th Ed., The C. V. Mosby Company, St. Louis, USA, 1980.
15. Kandel, E. R., Schwartz, J. H. and Jessel, T. M., (Eds.), Principles of Neural Science, 3rd Ed., Appleton & Lange, Norwalk, Connecticut, USA, 1991.
16. Farid, M., Murtagh, F., "Eye-movements and Voice as Interface Modalities to Computer Systems", SPIE – Opto Ireland, Galway, Ireland, 2002.
17. Farid, M., Murtagh, F., and Starck, JL., Computer Display Control And Interaction Using Eye-gaze, Journal of the Society for Information Display, 2002, in press.